|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continuous |
| Weight of Gold | Continuous |
| Distance between two places | Continuous |
| Length of a leaf | Continuous |
| Dog's weight | Continuous |
| Blue Color | Discrete |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

Q1) Identify the Data type for the Following:

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Ratio |
| Weight | Interval |
| Hair Color | Nominal |
| Socioeconomic Status | Nominal |
| Fahrenheit Temperature | Continuous |
| Height | Continuous |
| Type of living accommodation | Ordinal |
| Level of Agreement | Nominal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Ordinal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Interval |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained? **3/8 = 0.375**

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1 **0**
2. Less than or equal to 4 **6/36 = 1/6**
3. Sum is divisible by 2 and 3 **6/36 = 1/6**

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue? **10/21**

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

**Ans) Expected number of candies for a randomly selected child= 0.015+0.8+1.95+0.025+0.06+0.24 = 3.09**

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

**Ans)**

Mean for Points = 3.59, Score = 3.21 and Weigh = 17.84

Median for Points = 3.69, Score = 3.32 and Weigh = 17.71

Mode for Points = 3.07, Score = 3.44 and Weigh = 17.02

Variance for Points = 0.28, Score = 0.95, Weigh = 3.19

Standard Deviation for Points = 0.53, Score = 0.97, Weigh = 1.78

Range [Min-Max] for Points = [3.59 – 4.93], Score = [3.21 – 5.42] and

Weigh = [17.84 – 22.9]

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

**Ans)** Expected value = Sum (X \* Probability of X)= (1/9)(108)+ (1/9)(110)+ (1/9)(123)+ (1/9)(134)+ (1/9)(145)+ (1/9)(167)+ (1/9)(187)+ (1/9)(199)

= **145.33**

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

Skewness:

speed -0.117510

dist 0.806895

Kurtosis:

speed -0.508994

dist 0.405053

**SP and Weight(WT)**

**Use Q9\_b.csv**

Skewness:

SP 1.611450

WT -0.614753

Kurtosis:

SP 2.977329

WT 0.950291

**Q10) Draw inferences about the following boxplot & histogram**



**Ans)**  Skewness = 1.61 & Kurtosis = 0.95

🡨 Outliers

🡨 Upper Extreme UE = UQ + 1.5 IQR

🡨 Whiskers 1.5 IQR

🡨 Upper Quartile (IQR)

🡨 Median Inner Quartile Range

🡨 Lower Quartile

🡨 Lower Extreme 1.5 IQR

LE = LQ – 1.5 IQR



**Ans)** The boxplot has outliers beyond the maximum side Upper Extreme.

**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

**Ans)**

P = stats.t.interval(alpha = 0.94, df=1999, loc=200, scale=30/np.sqrt(2000))

np.round(P, 0)

print('CI for 94%:', P)

O/P:

CI for 94%: (198.7376089443071, 201.2623910556929)

CI for 98%: (198.4381860483216, 201.5618139516784)

CI for 96%: (198.6214037429732, 201.3785962570268)

**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?
3. Find mean, median, variance, standard deviation.

**Ans)** Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05

1. What can we say about the student marks?

**Ans)** We don’t have outliers and the data is slightly skewed towards right because mean is greater than median.

Q13) What is the nature of skewness when mean, median of data are equal? **Normal**

Q14) What is the nature of skewness when mean > median? **Negative**

Q15) What is the nature of skewness when median > mean? **Positive**

Q16) What does positive kurtosis value indicates for a data? **Positive Kurtosis means the curve will thinner and narrow / Narrow Peak, Smaller tail**

Q17) What does negative kurtosis value indicates for a data? **Negative Kurtosis means the curve will be flatter and broader / Wider Pear, Longer Tail**

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data? **Left tailed, Assymetrical, Median= 15, Upper Quartile = Upper Extreme= 18, Lower Extreme= 2, Outliers present**

What is nature of skewness of the data? **Negative**

What will be the IQR of the data (approximately)? 18-10= **8**

Q19) Comment on the below Boxplot visualizations?



**Ans)**

* There are no outliers.
* Both box plot shares the same median. (approx in a range between 275 to 250)
* 1) appears to be a bit right skewed(positive)
* 2) is normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

**Ans)**

**a)**

P\_MPG\_above\_38 = np.round(1 - stats.norm.cdf(38, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG>38) =', P\_MPG\_above\_38)

o/p: **P(MPG>38) = 0.348**

**b)** P\_MPG\_below\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<40) **=', P\_MPG\_below**\_40)

o/p: **P(MPG<40) = 0.729**

**c)** P\_MPG\_bw\_20\_50 = (np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) - np.round(1 - stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3))

print('P(20<MPG<50 =',P\_MPG\_bw\_20\_50)

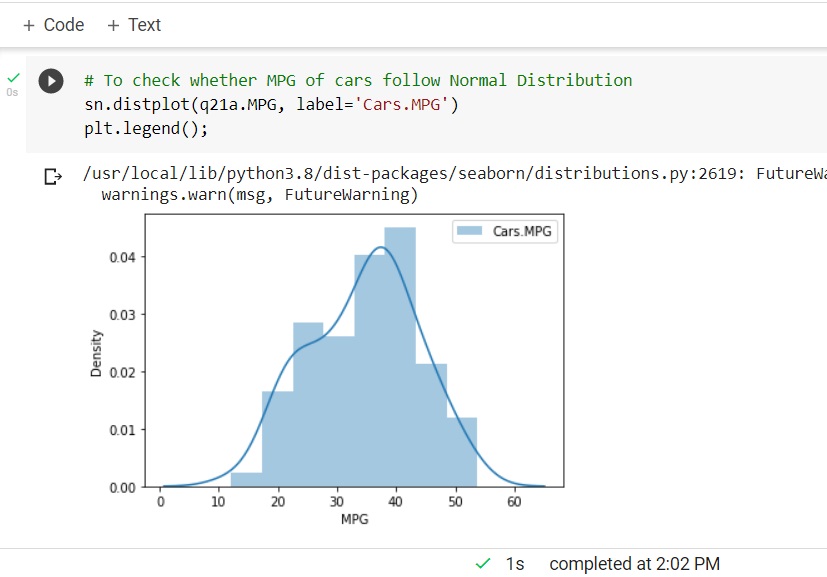
o/p: **P(20<MPG<50) = 0.013000000000000012**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans)**



1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

**Ans)**

plt.figure(figsize=(10,5))

plt.suptitle("Normal Distribution", fontsize=15)

plt.subplot(1,2,1)

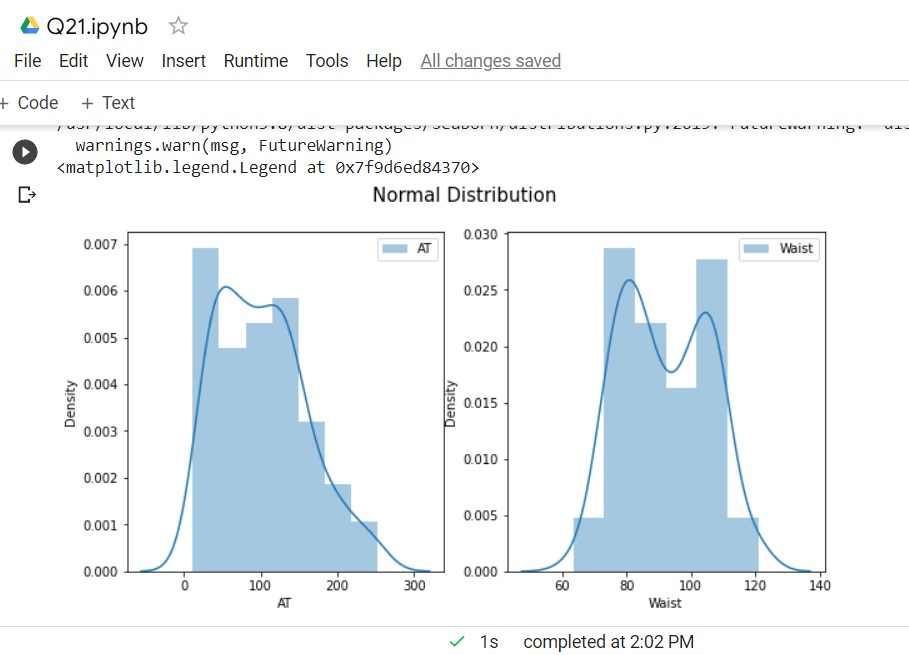
sn.distplot(q21b.AT, label="AT")

plt.legend()

plt.subplot(1,2,2)

sn.distplot(q21b.Waist, label="Waist")

plt.legend()



Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**Ans)**

print('Z-score for 90% Confidence Interval =', np.round(stats.norm.ppf(.05),4))

Z-score for 90% Confidence Interval = -1.6449

print('Z-score for 94% Confidence Interval =', np.round(stats.norm.ppf(.03),4))

Z-score for 94% Confidence Interval = -1.8808

print('Z-score for 60% Confidence Interval =', np.round(stats.norm.ppf(.2),4))

Z-score for 60% Confidence Interval = -0.8416

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

**Ans)**

print('T-score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4))

T-score for 95% Confidence Interval = -2.0639

print('T-score for 94% Confidence Interval =',np.round(stats.t.ppf(0.03,df=24),4))

T-score for 94% Confidence Interval = -1.974

print('T-score for 99% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4))

T-score for 99% Confidence Interval = -2.7969

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

**Ans)** μ=270, n=18, x=260, std=90,

# Null Hypothesis (Ho) = Avg life of Bulb >= 260 days

# Alternate Hypothesis (Ha) = Avg life of Bulb < 260 days

# find t-scores at x=260; t=(x-μ)/(std/sqrt(n))

t=(260-270)/(90/np.sqrt(18))

t

-0.4714045207910317

# Find Probability (p\_value) = 1-stats.t.cdf(abs(t\_scores),

df=n-1)... Using cdf function

p\_value = stats.t.cdf(t, df = 17)

print('Probability that 18 randomly selected bulbs would have an average life of no more than 260 days :', p\_value\*100,'%')

Probability that 18 randomly selected bulbs would have an average life of no more than 260 days : 32.16725356709836 %

Therefore,

Probability that 18 randomly selected bulbs would have an average life of no more than 260 days : **32.17 %**